网络大数据:现状与展望

王元卓 靳小龙 程学旗

(中国科学院计算技术研究所 网络数据科学与技术重点实验室 北京 100190)

摘 要 网络大数据是指"人、机、物"三元世界在网络空间(Cyberspace)中交互、融合所产生并在互联网上可获得的大数据.网络大数据的规模和复杂度的增长超出了硬件能力增长的摩尔定律,给现有的IT架构以及机器处理和计算能力带来了极大挑战.同时,也为人们深度挖掘和充分利用网络大数据的大价值带来了巨大机遇.因此,迫切需要探讨大数据的科学问题,发现网络大数据的共性规律,研究网络大数据定性、定量分析的基础理论与基本方法.文中分析了网络大数据的复杂性、不确定性和涌现性,总结了网络空间感知与数据表示、网络大数据存储与管理体系、网络大数据挖掘和社会计算以及网络数据平台系统与应用等方面的主要问题与研究现状,并对大数据科学、数据计算需要的新模式与新范式、新型的IT基础架构和数据的安全与隐私等方面的发展趋势进行了展望.

关键词 大数据;网络大数据;网络空间感知;大数据存储;数据挖掘;社会计算

中图法分类号 TP393 **DOI号** 10.3724/SP.J.1016.2013.01125

Network Big Data: Present and Future

WANG Yuan-Zhuo JIN Xiao-Long CHENG Xue-Qi

(*Key Laboratory of Web Data Science & Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

Abstract Network big data refer to the massive data generated by interaction and fusion of the ternary human-machine-thing universe in the Cyberspace and available on the Internet. The increase of their scale and complexity exceeds that of the capacity of hardware characterized by the Moore law, which brings grand challenges to the architecture and the processing and computing capacity of the contemporary IT systems, meanwhile presents unprecedented opportunities on deeply mining and taking full advantage of the big value of network big data. Therefore, it is pressing to investigate the disciplinary issues and discover the common laws of network big data, and further study the fundamental theory and basic approach to qualitatively or quantitatively dealing with network big data. This paper analyzes the challenges caused by the complexity, uncertainty and emergence of network big data, and summarizes major issues and research status of the awareness, representation, storage, management, mining, and social computing of network big data, as well as network data platforms and applications. It also looks ahead to the development trends of big data science, new modes and paradigm of data computing, new IT infrastructures, and data security and privacy, etc.

Keywords big data; network big data; cyberspace awareness; storage of big data; data mining; social computing

---

收稿日期:2012-12-18;最终修改稿收到日期:2013-03-20.本课题得到国家自然基金重点项目"在线社会关系网络挖掘与分析"(61232010)、"支持舆情监控的Web搜索与挖掘的新理论和新方法"(60933005)、国家"九七三"重点基础研究发展规划项目课题"面向公共安全的社会感知数据处理"(2012CB316303);国家自然基金面上项目"基于随机博弈网的网络用户信息行为模型及演化性分析"(61173008)、国家自然科学基金青年项目"通信网络中可变服务容量调度系统的性能建模、分析与优化"(61100175)资助.王元卓,男,1978年生,博士,副研究员,中国计算机学会(CCF)高级会员,主要研究方向为社会计算、网络行为分析、信息安全等.E-mail:wangyuanzhuo@ict.ac.cn.靳小龙,男,1976年生,博士,副研究员,主要研究方向为社会计算、网络性能建模与分析、多智能体系统等.程学旗,男,1971年生,博士,研究员,主要研究领域为网络科学、网络与信息安全以及互联网搜索与服务.

1 引 言

1.1 研究与发展现状

近年来,随着互联网、物联网、云计算、三网融合等 IT 与通信技术的迅猛发展,数据的快速增长成了许多行业共同面对的严峻挑战和宝贵机遇,因而信息社会已经进入了大数据(Big Data)时代.大数据的涌现不仅改变着人们的生活与工作方式、企业的运作模式,甚至还引起科学研究模式的根本性改变.

一般意义上,大数据是指无法在一定时间内用常规机器和软硬件工具对其进行感知、获取、管理、处理和服务的数据集合[1].网络大数据是指"人、机、物"三元世界在网络空间(Cyberspace)中彼此交互与融合所产生并在互联网上可获得的大数据,简称网络数据.

当前,网络大数据在规模与复杂度上的快速增长对现有 IT 架构的处理和计算能力提出了挑战.据著名咨询公司 IDC 发布的研究报告,2011 年网络大数据总量为 1.8 ZB,预计到 2020 年,总量将达到 35 ZB.

IBM 将大数据的特点总结为 3 个 V,即大量化(Volume)、多样化(Variety)和快速化(Velocity).首先,网络空间中数据的体量不断扩大,数据集合的规模已经从 GB、TB 到了 PB,而网络大数据甚至以 EB 和 ZB(10^{21})等单位来计数.IDC 的研究报告称,未来十年全球大数据将增加 50 倍,管理数据仓库的服务器的数量将增加 10 倍以迎合 50 倍的大数据增长①.其次,网络大数据类型繁多,包括结构化数据、半结构化数据和非结构化数据.在现代互联网应用中,呈现出非结构化数据大幅增长的特点,至 2012 年末非结构化数据占有比例达到互联网整个数据量的 75%以上.这些非结构化数据的产生往往伴随着社交网络、移动计算和传感器等新技术的不断涌现和应用.再次,网络大数据往往呈现出突发涌现等非线性状态演变现象,因此难以对其变化进行有效评估和预测.另一方面,网络大数据常常以数据流的形式动态、快速地产生,具有很强的时效性,用户只有把握好对数据流的掌控才能充分利用这些数据.

近几年,网络大数据越来越显示出巨大的影响作用,正在改变着人们的工作与生活.2012 年 11 月《时代》杂志撰文指出奥巴马总统连任成功背后的秘密,其中的关键是对过去两年来相关网络数据的搜集、分析和挖掘②.目前,eBay 的分析平台每天处理的数据量高达 100 PB,超过了纳斯达克交易所每天的数据处理量.为了准确分析用户的购物行为,eBay 定义了超过 500 种类型的数据,对顾客的行为进行跟踪分析③.2012 年的双十一,中国互联网再次发生了最大规模的商业活动:淘宝系网站的销售总额达到 191 亿元人民币.淘宝之所以能应对如此巨大的交易量和超高并发性的分析需求,得益于其对往年的情况,特别是用户的消费习惯、搜索习惯以及浏览习惯等数据所进行的综合分析④.

网络大数据给学术界也同样带来了巨大的挑战和机遇.网络数据科学与技术作为信息科学、社会科学、网络科学和系统科学等相关领域交叉的新兴学科方向正逐步成为学术研究的新热点.近年,《Nature》和《Science》等刊物相继出版专刊来探讨对大数据的研究.2008 年《Nature》出版的专刊"Big Data",从互联网技术、网络经济学、超级计算、环境科学和生物医药等多个方面介绍了海量数据带来的挑战[2].2011 年《Science》推出关于数据处理的专刊"Dealing with Data",讨论了数据洪流(Data Deluge)所带来的机遇[3].特别指出,倘若能够更有效地组织和使用这些数据,人们将得到更多的机会发挥科学技术对社会发展的巨大推动作用.

1.2 网络大数据研究的意义

总体而言,网络大数据研究的重要性体现在以下几个方面:

(1)网络大数据的研究对捍卫国家网络空间的数字主权,维护社会稳定,推动社会与经济可持续发展有着独特的作用.信息化时代,国家层面的竞争力将部分体现为一国拥有网络大数据的规模、活性以及对数据的解释与运用的能力.国家在网络空间的数字主权也将是继海、陆、空、天四空间之后另一个大国博弈的空间.在网络大数据领域的落后,意味着失守产业战略制高点,意味着国家安全将在网络空间出现漏洞.为此,今年 3 月,美国政府整合 6 个部门投资 2 亿美元启动"大数据研究和发展计划".在该计划中,美国国家科学基金会提出要"形成一个包括数学、统计基础和计算机算法的独特学科".该计划还强调,大数据技术事关美国的国家安全,影响科学研究的步伐,还将引发教育和学习的变革.这意味

① http://www.emc.com/
② http://swampland.time.com//
③ http://www.china-cloud.com/
④ http://server.51cto.com/

着网络大数据的主权已上升为国家意志，直接影响国家和社会的稳定，事关国家的战略安全.

(2) 网络大数据是国民经济核心产业信息化升级的重要推动力量. “人、机、物”三元世界的融合产生了大规模的数据，如何感知、测量、利用这些网络大数据成为国民经济中许多行业面临的共同难题，成为这些行业数字化、信息化的障碍和藩篱. 如何使不同行业都能突破这一障碍，关键在于对网络大数据基本共性问题的解决. 譬如，对于非结构化数据的统一表示与分析，目前缺少有效的方法和工具. 因此，通过对网络大数据共性问题的分析和研究，使企业能够掌握网络大数据的处理能力或者能够承受网络大数据处理的成本与代价，进而使整个行业迈入数字化与信息化的新阶段. 在这个意义上，对网络大数据基础共性问题的解决将是新一代信息技术融合应用的新焦点，是信息产业持续高速增长的新引擎，也是行业用户提升竞争能力的新动力.

(3) 网络大数据在科学和技术上的突破，将可能诞生出数据服务、数据材料、数据制药等战略性新兴产业. 网络数据科学与技术的突破意味着人们能够理清数据交互连接产生的复杂性，掌握数据冗余与缺失双重特征引起的不确定性，驾驭数据的高速增长与交叉互连引起的涌现性(Emergence)[4]，进而能够根据实际需求从网络数据中挖掘出其所蕴含的信息、知识甚至是智慧，最终达到充分利用网络数据价值的目的. 涌现性是指由低层次的多个元素构成高层次的系统时展示出的每个单一元素所不具备的性质. 网络数据不再是产业环节上产生的副产品，相反地，网络数据已成为联系各个环节的关键纽带. 通过对网络数据纽带的分析与掌握，可以降低行业成本、促进行业效率、提升行业生产力. 因此，可以预见，在网络数据的驱动下，行业模式的革新将可能催生出数据材料、数据制造、数据能源、数据制药等一系列战略性的新兴产业.

(4) 大数据引起了学术界对科学研究方法论的重新审视，正在引发科学研究思维与方法的一场革命. 科学研究最初只有实验科学，随后出现了理论科学，研究各种定律和定理. 由于在许多问题上，理论分析方法变得太过复杂以至于难以解决难题，人们开始寻求模拟的方法，这又产生了计算科学. 而大数据的出现催生了一种新的科研模式，即面对大数据，科研人员只需从数据中直接查找、分析或挖掘所需要的信息、知识和智慧，甚至无需直接接触需研究的对象. 2007 年，已故的图灵奖得主吉姆格雷(Jim Gray)在他最后一次演讲中描绘了数据密集型科学研究的“第四范式”(The Fourth Paradigm)[5]，把数据密集型科学从计算科学中单独区分开来. 格雷认为，要解决我们面临的某些最棘手的全球性挑战，“第四范式”可能是唯一具有系统性的方法.

网络大数据的深挖掘、大规模利用是新兴产业界的立足点. 即便针对大数据的研究目前还没有建立一套完整的理论体系，也缺少高效快速的处理、分析与挖掘的算法与范式，但大数据的应用前景毋庸置疑，因为大数据从根本上来说就是来源于应用的问题. 著名出版公司 O’Reilly 的创始人 Tim O’Reilly 断言，大数据就是下一个 Intel Inside，未来属于那些能把数据转换为产品的公司和人群. MGI 的研究报告也宣称，大数据是下一代革新、竞争力和生产力的先导，网络大数据可为世界经济创造巨大价值，提高企业和公共部门的生产率和竞争力，并为消费者创造巨大的经济利益. Gartner 公司则更具体地预测，到 2015 年，采用大数据和海量信息管理的公司将在各项财务指标上，超过未做准备的竞争对手 20%.

本文梳理了网络大数据所带来的挑战以及相关的研究体系，从网络空间感知与数据表示、网络大数据存储与管理体系、网络数据挖掘和社会计算以及网络数据平台系统与应用 4 个方面回顾了相关领域的新近发展，探讨了网络大数据研究方向和所面临的挑战，并展望了未来的主要研究方向.

2　网络大数据带来的挑战

如上所述，网络大数据面临着来自诸多方面的挑战. 但从研究的角度来说，根本挑战在于其复杂性、不确定性和涌现性. 对这 3 个基本特性的研究决定着网络大数据的发展趋势、研究进展和应用前景.

2.1　网络大数据的复杂性

复杂性造成网络大数据存储、分析、挖掘等多个环节的困难. 网络大数据的复杂性主要包括数据类型的复杂性、数据结构的复杂性和数据内在模式的复杂性.

(1) 数据类型复杂性. 信息技术的发展使得数据产生的途径不断增加，数据类型持续增多. 相应地，则需要开发新的数据采集、存储与处理技术. 例如社交网络的兴起，使得微博、SNS 个人状态信息等短文本数据逐渐成为互联网上的主要信息传播媒介. 与传统的长文本不同，短文本由于长度短，上下

文信息和统计信息很少，给传统的文本挖掘（如检索、主题发现、语义和情感分析等）带来很大的困难. 相关的研究包括利用外部数据源（如 Wikipedia[6]、搜索结果[7]等）扩充文档，或者利用内部相似文档信息来扩充短文本的表达[8]. 然而，无论是利用外部数据，还是利用内部数据，都可能引入更多的噪声. 另一方面，不同数据类型的有机融合给传统的数据处理方法带来了新的挑战. 例如在社交媒体的研究当中地域信息与内容的融合[9]、时空信息与内容信息的结合[10]等等.

（2）数据结构的复杂性. 传统上处理的数据对象都是有结构的，能够存储到关系数据库中. 但随着数据生成方式的多样化，如社交网络、移动计算和传感器等技术，非结构化数据成为大数据的主流形式. 非结构化数据具有许多格式，包括文本、文档、图形、视频等等. 非结构化数据当中蕴含着丰富的知识，但其异构和可变的性质也给数据分析与挖掘工作带来了更大的挑战. 与结构化的数据相比，非结构化数据相对组织凌乱，包含更多的无用信息，给数据的存储与分析带来很大的困难. 目前相关的研究热点，包括开发非关系型数据库（如 Google 的 BigTable，开源的 HBase 等）来存储非结构化数据. Google 提出了 MapReduce 计算框架，Yahoo !、Facebook 等公司在此基础上实现了 Hadoop、Hive 之类的分布式架构，对非结构化数据做基本的分析工作. 国内各大公司和科研单位也启动了用于支撑非结构化处理的基础设施研发，如百度的云计算平台、中国科学院计算技术研究所的凌云（LingCloud）系统等.

（3）数据模式的复杂性. 随着数据规模的增大，描述和刻画数据的特征必然随之增大，而由其组成的数据内在模式将会以指数形式增长. 首先，数据类型的多样化决定了数据模式的多样化. 不仅需要熟悉各种类型的数据模式，同时也要善于把握它们之间的相互作用. 这种面向多模式学习的研究需要综合利用各个方面的知识（如文本挖掘、图像处理、信息网络、甚至社会学等等）. 为此，Sun 提出用网络来描述异质数据间的关系，同时提出了“元路径（Meta-Path）”的概率来刻画目标数据模式[11]. 这样，通过定义合适的元路径，便可在数据网络中挖掘有价值的模式. 其次，非结构化的数据通常比结构化数据蕴含更多的无用信息和噪声，网络数据需要高效鲁棒的方法来实现去粗存精，去冗存真. 搜索引擎就是从无结构化数据中检索出有用信息的一种工具. 尽管搜索技术在工业上已经取得极大的成功，但仍然存在很多不足（如对一些长尾词的查询，对二义性查询词的理解等），都有待进一步提高. 另外，网络大数据通常是高维的，往往会带来数据高度稀疏与维度灾难等问题. 在这种情况下，由于数据模式统计显著性较弱，以往的统计学习方法多针对高频数据挖掘模式，因此难以产生令人满意的效果. 近年来，受实际应用驱动，高维稀疏问题成为了统计学习领域的热点问题[12]. 相关理论研究发现，基于稀疏表达的学习方法（如 LASSO 等），在获得较好学习效果的同时，还具有更高的效率和鲁棒性[13].

2.2 网络大数据的不确定性

不确定性使得网络数据难以被建模和学习，从而难以有效利用其价值. 网络数据的不确定性包括数据本身的不确定性、模型的不确定性和学习的不确定性.

（1）数据的不确定性. 原始数据的不准确以及数据采集处理粒度、应用需求与数据集成和展示等因素使得数据在不同维度、不同尺度上都有不同程度的不确定性. 传统侧重于准确性数据的处理方法，难以应对海量、高维、多类型的不确定性数据. 具体而言，在数据的采集、存储、建模、查询、检索、挖掘等方面都需要有新的方法来应对不确定性的挑战[14]. 近年来，概率统计的方法被逐步应用于不确定性数据的处理中. 一方面，数据的不确定性要求我们使用不确定的方法加以应对；另一方面，计算机硬件的发展也为这类方法提供了效率、效能上的可能. 目前，该领域研究尚浅，在学术界和产业界尚有大量问题亟待解决.

（2）模型的不确定性. 数据的不确定性要求对数据的处理方式能够提出新的模型方法，并能够把握模型的表达能力与复杂程度之间的平衡. 在对不确定数据的建模和系统设计上，最常用且朴素的观点是“可能世界模型”[15]. 该观点认为，在一定的结构规范下，应将数据的每一种状态都加以刻画. 但该种模型过于复杂，难以用一种通用的模型结构来适应具体的应用需求. 在实际应用中，我们往往采取简化的模型刻画不确定性数据的特性，如独立性假设、同分布假设等等. 尤其值得注意的是，概率图模型[16]由于具有很强的表达能力而且可对数据相关性进行建模，因此已被广泛应用在不确定数据的建模领域. 另外，在数据的管理和挖掘上面，不确定性模型的构建应当考虑到数据的查询、检索、传输、展示等方面的影响[17].

（3）学习的不确定性. 数据模型通常都需要对

模型参数进行学习. 然而，在很多情况下找到模型的最优解是 NP 问题，甚至找到一个局部最优解都很困难. 因此很多学习问题都采用近似的、不确定的方法来寻找一个相对不错的解. 但在大数据的背景下，传统近似的、不确定的学习方法需要面对规模和时效的挑战. 随着多核 CPU/GPU 的普及以及并行计算框架的研究，分而治之的方法被普遍认为是解决网络大数据问题一条必由之路. 如何将近似的、不确定的学习方法拓展到这种框架上成为当前研究的重点. 近年来，不少高校和研究机构，在该领域做出了探索. 如在矩阵分解运算中对数据进行分块的计算方法能够利用多台机器并行计算，从而提高数据的处理速度[18-19]. 此外，除了学习模型参数值的不确定外，模型的复杂性和参数个数也受到不同领域、不同数据类型和应用需求的影响而不能提前确定. 近年来，在统计学习领域，非参模型方法的提出[20-21]为自动学习出模型复杂度和参数个数提供了一种思路. 但该类模型计算上较为复杂，如何分布式地、并行地应用到网络大数据的处理上，还是一个开放问题.

2.3　网络大数据的涌现性

涌现性是网络数据有别于其它数据的关键特性. 涌现性在度量、研判与预测上的困难使得网络数据难以被驾驭. 网络数据的涌现性主要表现为模式的涌现性、行为的涌现性和智慧的涌现性.

(1) 模式的涌现性. 在多尺度、异质关系的网络数据中，由于不同的数据在属性、功能等方面既存在差异又相互关联，因此使网络大数据在结构、功能等方面涌现出了局部结构所不具备的特定模式特征. 在结构方面，数据之间不同的关联程度使得数据构成的网络涌现出模块结构. 在功能方面，网络在演化过程中会自发地形成相互分离的连通小块[22-24]. 这一涌现性结果对于研究更多的社会网络模型和理解网络瓦解失效的发生有着重要意义.

(2) 行为的涌现性. 随着数据采集技术的不断发展，人们得到的很多数据都具有时序性，而社会网络中个体行为的涌现性则是基于数据时序分布的统计结果. 在社会网络中有较大相似性的个体之间容易建立社会关系. 通过研究 Schelling 给出的个体社会关系网络模型发现，网络在演化过程中会自发地形成相互分离的连通块，这一个体行为涌现的结果不依赖于初始网络的拓扑结构，对于研究更多的社会网络模型和理解行为涌现的规律具有重要意义[25]. 著名网络科学家 Barabasi 研究发现，人们发邮件的数量在一天的某些时刻会出现“爆发”现象，并发现每个人连发两封邮件之间的时间间隔涌现出幂率分布特征. 此外，自然界和社会中个体之间不同的竞争模式会导致不同的同步状态的涌现性.

(3) 智慧的涌现性. 网络数据在没有全局控制和预先定义的情况下，通过对来自大量自发个体的语义进行互相融合和连接而形成语义，整个过程随着数据的变化而持续演进，从而形成网络数据的涌现语义，也可以称之为智慧涌现. 作为一种特殊的智慧涌现形式，众包正在通过互联网和社会网络快速发展，成为一种新的商业模式、新的数据产生模式和新的数据处理协作模式.

总体而言，尽管与网络大数据研究密切相关的数据库、数据挖掘、机器学习和知识工程等领域近些年来都有很大的进展，甚至在许多不同的领域得到了深入的应用，但由于网络大数据规模海量、关系复杂等根本特性，使得相关领域的研究成果难以被直接借鉴于网络大数据的研究. 因此，网络大数据的研究需要一套全新的理论和方法来进行方向性的指导. 但到目前为止，甚至连大数据的精确定义还缺乏一个统一的标准. 网络大数据科学与技术这门学科的内涵和外延还缺乏严格的限定和详实的论证；在大数据的环境下，传统“假设、模型、检验”的科学方法受到质疑，从“数据”到“数据”的第四范式还没有建立，需要一个完备的新的理论体系来指导该学科的发展和研究.

3　网络空间感知与数据表示

网络数据具有跨媒体关联、强时效演变、多主体互动等特点，使得我们对网络大数据的态势感知、质量评估、融合表示等均面临新的问题.

3.1　网络大数据的感知与获取

按照网络空间中数据的蕴藏深度，整个网络空间可以划分为 Surface Web 和 Deep Web[26]，或称作 Hidden Web[27]. Surface Web 是指 Web 中通过超链接可被传统搜索引擎爬取到的静态页面，而 Deep Web 则由 Web 中可在线访问的数据库组成. Deep Web 的数据隐藏在 Web 数据库提供的查询接口后面，只有通过向查询接口提交查询才能获得. 与 Surface Web 相比，Deep Web 所包含的信息更丰富. 同时，Deep Web 具有规模大、实时动态变化、异构性、分布性以及访问方式特殊等特点. 为了充分利用 Deep Web 中的数据资源，需要充分获取 Deep Web 中高质量的数据并予以集成，整个集成过程可

以分为数据获取、数据抽取和数据整合 3 个环节.

3.2 网络大数据的质量评估与采样

对网络空间中多源数据进行质量评估,一方面需要建立数据模型或提出适当的采样方法;另一方面,需要提出对采样数据的评价与检验方法.网络数据采样是将数据从 Web 数据库提取出来的过程.传统的数据库采样是随机从数据库中选取数据记录以获得数据库的统计信息的过程,典型方法可参考文献[28-30].但是要获取 Web 数据库中的数据只能通过向查询接口提交查询,不能自由地从 Web 数据库获取记录,故而传统方法不能实现对 Web 数据库的采样.

针对 Web 数据库采样,HIDDEN-DB-SAMPLER[31]是第一项工作,它给出了对范围属性和分类属性的处理方法,而对查询接口中设计的必填的可任意取值的关键词属性未作处理.文献[32]提出基于图模型的增量式 Web 数据库采样方法 WDB-Sampler,通过查询接口从 Web 数据库中以增量的方式获取近似随机的样本.但是该方法是针对样本中每条数据作为顶点来建立图模型,每一轮查询后都要将查询结果扩充到图模型中用于产生下一轮查询词,这样做的代价比较高.

3.3 网络大数据的清洗与提炼

由于现实世界数据的多源性、异质性以及采集数据时的一些人工错误,导致网络数据是含有噪音、冗余和缺失的.如何有效地衡量数据的质量是一个重要的研究方向.文献[33]定义了衡量数据质量的 4 个指标:一致性、正确性、完整性和最小性.文献[34]提出了数据工程中数据质量的需求分析和模型,认为存在很多候选的数据质量衡量指标,用户应根据应用的需求选择其中一部分.

数据的清洗建立在数据质量标准之上,为了得到高质量的数据,清洗与提炼过程必须满足几个条件:检测并除去数据中所有明显的错误和不一致;尽可能地减小人工干预和用户的编程工作量,而且要容易扩展到其它数据源;应该和数据转化相结合;要有相应的描述语言来指定数据转化和数据清洗操作,所有这些操作应该在一个统一的框架下完成.对于数据清洗,工业界已经开发了很多数据抽取、转化和装载工具(ETL tool)[35].一些研究人员研究相似重复记录的识别和剔除(如文献[34,36]),还有一些研究包括数据的变换和集成(如文献[37-38]).

3.4 网络大数据的融合表示

对网络数据的建模和表达理论方面的研究,主要集中在网络中的文本信息方面.对文本信息进行表示和建模其目的是让计算机能够正确理解人类的语言,能够分析和表达出其中的语义信息.文本信息的表达经历了从浅层词语表达方式到深层语义表达方式这样一个历程,其中代表性的工作包括了向量空间表示(VSM)[29]、隐语义索引(LSI)[39]和概率话题模型(如图 1 所示)[40]等.随着研究不断深入,话题模型被广泛地应用在各个领域,进一步有人提出了改进的话题模型[41],以增强已有话题模型的学习能力,解决其跨领域的问题等等,从而使其能更好地应用于文本数据的表达.

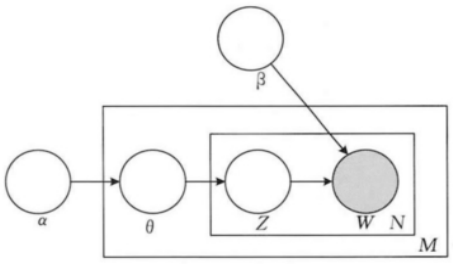

图 1 概率话题模型[40]

尽管对数据表达的研究历经了很长的时间,但是对于网络大数据的建模和表达还面临着很多新的挑战.例如,对于海量文本数据的建模,我们需要模型能够对更大规模的参数空间进行有效地学习,需要能够有效地建模并解决数据的稀疏性所带来的问题,需要能够对动态演化的网络大数据进行合理的表达.此外,对于图片和多媒体数据,我们也需要进一步探索其建模与表达方式,以便能够更加有效地表达其内在的语义信息.

4 网络大数据存储与管理体系

网络大数据处理的数据规模从 TB 级上升到 PB、EB 级,面临着如何降低数据存储成本、充分利用计算资源、提高系统并发吞吐率、支持分布式的非线性迭代算法优化等众多难题.

4.1 分布式数据存储

Google 公司提出的 GFS、MapReduce、BigTable 等技术是分布式数据处理技术的具体实现,是 Google 搜索引擎系统三大核心技术.此后,Apache 软件基金会推出了开放源码 Hadoop 和 HBase 系统,实现了 MapReduce 编程模型、分布式文件系统和分布式列簇数据库.Hadoop 系统在 Yahoo !、IBM、百度、Facebook 等公司得到了大量应用和快速的发展.但作为一个新兴的技术体系,分布式数据

处理技术在支持大规模网络信息处理及应用等大数据计算应用方面还存在着很多不足.

行存储(Row-Store)和列存储(Column-Store)是两种典型的数据库物理存储策略.行存储方式较为传统,它在磁盘中依次保存每条记录,比较适合事务操作;列存储方式垂直划分关系表,以列为单位存储数据,列存储还具有数据压缩(Compression)、延期物化(Late Materialization)、块循环(Block Iteration)等特性[42].由于数据分析任务往往仅使用较少字段,因此列存储方式的效率更高.数据分析任务在大数据应用中更为常见,因此许多系统尽管无法完全实现列存储的所有特性,但也或多或少地借鉴了相关概念,包括 BigTable、HBase 等[43].文献[44]提出了行列混合式数据存储结构(RCFile)以解决海量数据快速加载、缩短查询响应时间、磁盘空间高效利用等问题(如图 2 所示).RCFile 融合了行存储和列存储的优点,通过行组划分降低数据加载开销,通过列数据压缩提高存储空间利用率.国际上应用最广泛的两大分布式数据分析系统 Hive 和 Pig 均集成了 RCFile 技术.RCFile 已经成为分布式离线数据分析系统中数据存储结构的事实标准.

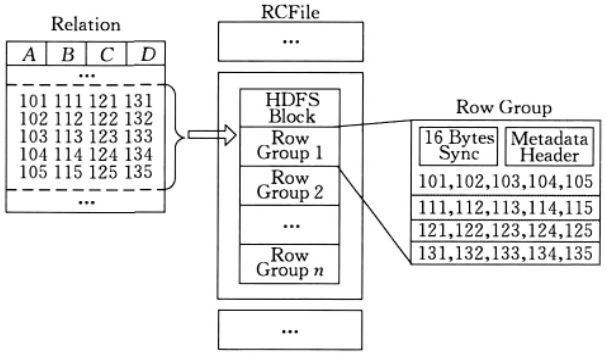

图 2　RCFile 数据存储结构示例[44]

分布式数据存储是网络大数据应用的一个重要环节.但目前的研究工作仍存在一些局限性.针对海量数据存储和处理所面临的数据总量超大规模、处理速度要求高和数据类型异质多样等难题,需要开发支持高可扩展、深度处理的 PB 级以上分布式数据存储框架,同时需要研究适应数据布局分布的存储结构优化方法,以提高网络大数据存储和处理效率,降低系统建设成本,从而实现高效、高可用的网络大数据分布式存储.

4.2　数据高效索引

目前的主流查询索引技术是以 Google 公司的 BigTable 为代表的列簇式 NoSQL 数据库.BigTable 提出了一种介于关系模型和 Key-Value 对模型之间的新数据模型:Ordered Table.Ordered Table 模型提供了稀疏的、分布式的、持久存储的、基于主键排序的映射,数据由行、列和时间戳表示.BigTable 中表的 Scheme 非常灵活,可以在运行时修改.Ordered Table 模型可以对基于主键的区间查询提供有力的支持,对于涉及多个字段数据的多维区间查询主要采用二级索引技术,但这引起了性能问题.

为避免大量随机读,另一种思路是使用聚簇索引,即同时按索引顺序存储全部数据.为保证多个查询列均有很好的性能,需要按多个索引列聚簇,但这将导致存储开销成倍增长.此外还面临着因统计信息的缺失带来的新的挑战.关系数据库领域处理多个维度的查询优化时,关键是根据表的统计信息估算子查询的代价,比如通过记录数量、数据分布的直方图等估算结果集大小、需要读取的数据块数量等.

文献[45]提出的互补式聚簇索引(CCIndex),利用多副本为每个索引列各创建一张互为补充的聚簇索引表,使得索引列上的区间查询对应聚簇索引表的连续扫描(如图 3 所示).解决了 NoSQL 数据库的二级索引技术因无法保持连续扫描特性而处理效率低下的问题.同时,结合查询结果集估算方法,以挑选最优查询计划.通过禁用底层存储系统的副本机制来避免引入额外的存储开销,并提供增量式的快速数据恢复机制.目前,CCIndex 技术已应用在淘宝的数据魔方中.

4.3　数据世系管理

数据世系(Data Provenance)[46]包含了不同数据源间的数据演化过程和相同数据源内部数据的演化过程.数据世系一般有两类基本方法,非注解的方法和基于注解的方法.前者采用模式映射方式使用数据处理函数和其相对应的反向函数,但在更复杂的例子中可能并不存在集合之间的可逆函数,必须使用注解描述世系.事实上基于注解的方法的应用范围要远远高于非注解的方法.

数据世系可针对多种数据类型,包括关系型数据、XML 数据和不确定数据等.自 20 世纪 90 年代以来,数据世系的研究取得很大的进展[47-48],并且应用到了多个领域之中.面对网络大数据,数据世系管理的研究工作需关注以下几个方面[49]:(1)传统的数据管理下的数据世系的管理还有很多的工作亟待考虑,其中考察数据的起源和演化过程将是一个大的挑战;(2)在网络环境下不确定性数据广泛存在,

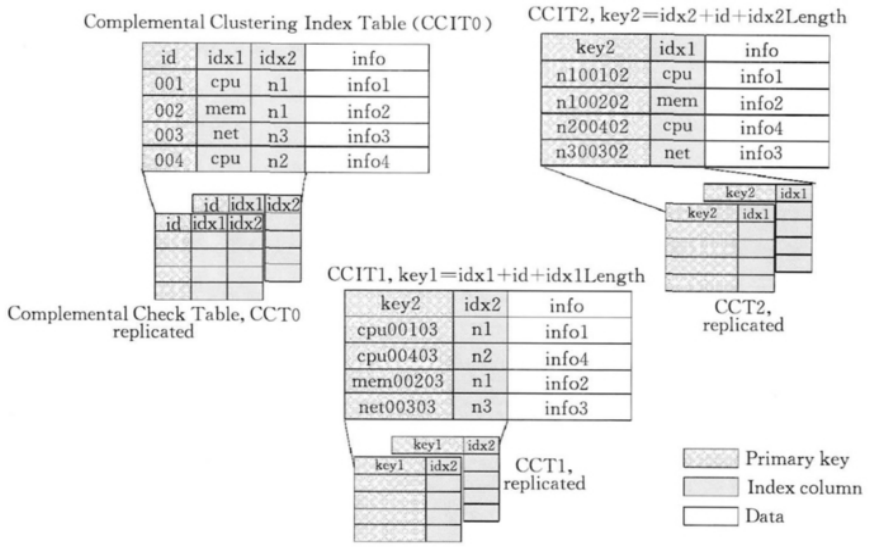

图 3 互补聚簇索引表[45]

并且具有多种多样的表现形式. 数据的演化过程同时也伴随着数据不确定性的演化,可以利用数据的世系追踪数据不确定性的来源和演化过程;(3)如何解决异构世系标准的融合问题. 大数据应用将涵盖更多的原本可能相互隔离的数据集合,如何将适用不同标准的数据世系信息整合在一起是一个关键问题.

5 网络大数据挖掘和社会计算

利用计算技术对网络大数据进行挖掘分析,发现蕴含的知识,研究社会运行的规律与发展趋势,是挖掘网络大数据的深层价值和实现社会行为可计算的主要途径. 随着社会媒体的涌现,持续增长的用户数据在规模和复杂性上都有着指数式的攀升,导致传统的挖掘和计算方法在性能和效用上遇到了严重的瓶颈. 基于内容信息的数据挖掘和基于结构信息的社会计算是目前网络大数据挖掘和社会计算领域的研究热点.

5.1 基于内容信息的数据挖掘

语言是社会媒体最重要的表现形式,文本是社会媒体中用户表达信息的最重要的方式. 基于内容信息的数据挖掘包括网络搜索技术与实体关联分析等主要研究内容.

社会媒体的出现为互联网信息搜索提出了新的挑战,研究的热点从传统的海量数据抓取、索引结构优化和用户查询分析等转移到了排序学习算法,专注于提高检索质量. 排序学习模型将文档表示为特征向量,以损失函数为优化目标,寻找在检索领域中常用的评价准则下最好的排序函数,常见的排序学习算法可以分为逐点(Pointwise,如 McRank[50])、逐对(Pairwise,如 RankBoost[51]、RankNet[52])和逐列(Listwise,如 ListNet[53]、AdaRank[54]、SVM-MAP[55])3 类方法. 现有模型在处理用户需求相关性、多样性和重要性等不同目标排序方面仍有不足. 此外,社会媒体中需要关注数据的短文本特征、对简短关键词表达的深入理解和分析,掌握用户真实的查询意图[56].

命名实体是现实世界中的具体或者抽象但具有特定意义的实体,从海量信息中获取其蕴含的内在知识,需要研究对命名实体、实体关系的挖掘. 社会媒体生成的海量网络数据中,实体类型越来越多,力度越来越细,关系越来越繁杂. 对于实体关系的挖掘,研究人员提出了基于规则[57]和基于机器学习[58]的方法. 2007 年,Getoor 等提出统计关系学习是里程碑式的技术[59],突破了传统统计模型对于研究对象同类型、不相关的两个假设,可以更全面地表达领域知识. 目前,实体和关系的挖掘仍是网络数据挖掘领域关注的研究问题,存在很多亟待解决的问题,例如对新涌现出的实体的抽取与识别,挖掘结果的可用性和可理解性,大规模高效知识库、本体库语义网络的构建等.

5.2 基于结构信息的社会计算

社会网络是以社会媒体中的用户为节点,用户间的关系为连边而构建的网络. 它既是用户间社会关系的反映,也是用户之间进行信息交互的载体. 具

有关系的异质性、结构的多尺度性以及网络的动态演化性 3 方面特性. 社会网络中个体因血缘关系或兴趣爱好等因素而形成了连接紧密的圈子，这种内部关系紧密而对外关系相对稀疏的结构被称为社区. 社区结构是社会网络所普遍具有的结构特征，社区结构的存在对于网络的高效搜索、网络演化、信息扩散等具有重要意义. 针对社区结构的研究可分为社区发现、社区结构演化等方面[60].

社区发现[61]旨在识别出网络固有的社区结构，按照节点间的连边关系把节点划分成若干节点组，使得节点内部的连边相对稠密，不同节点之间的连边相对稀疏. Girvan 和 Newman[62]提出分裂式层次聚类方法，是一种自顶向下的社区分割过程；文献[63]提出模块度概念，采用一种被假定没有社区结构的网络作为参照网络，对于一个给定的网络划分，通过对比原有网络和参照网络中处于该划分的各个分量内部边的比例，给出一种度量网络划分质量的方法；对于重叠社区结构的研究，Palla 等人[64]提出了一种基于完全子图渗流的社区发现方法，已应用到生物、信息、社会等网络中；进一步，文献[65]定义新的网络模块度，采用聚合式层次聚类的方式，提出了能够同时揭示网络层次重叠社区结构（如图 4 所示）的社区发现方法.

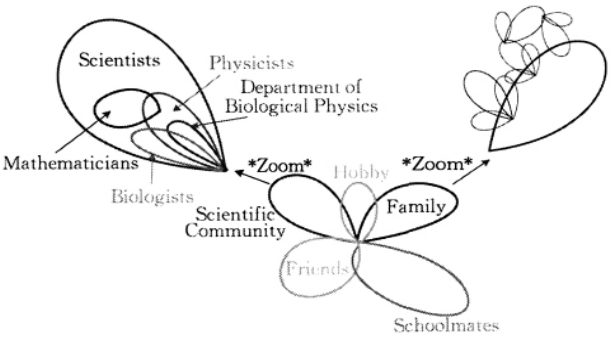

图 4　层次重叠社区结构示意图[65]

社区演化是网络自身结构与在其上频繁发生的交互过程相互作用的结果. 社区演化分析主要研究社区随时间变化的情况，并分析导致这些变化的机制和原因，包括社区的形成、生长、缩减、合并、分裂和消亡等. 在动态演化过程网络建模研究方面，Barabási 和 Albert[66]提出了著名的 BA 网络生成模型，建立了网络微观机制和宏观拓扑结构特征的关联规律. 文献[67]基于完全子图渗流社区发现方法研究社区演化，得出了小社区稳定性是保证其存在的前提而大社区的动态性是存在的基础的结论. 随着含时间数据的积累，关于社区演化的研究将会是一个热点.

6　网络数据平台系统与应用

为了应对网络大数据的发展趋势，更好地为企业和个人提供数据分析的需求，亟需构建各类不同的网络大数据平台，支持用户对数据的多种需求. 下面我们从数据平台建设、基于数据平台的高端数据分析以及网络大数据平台的应用 3 个角度总结相关的内容.

6.1　网络大数据平台引擎建设

构建网络大数据平台就是要将不同渠道、不同来源、不同结构的数据进行有机的整合. 与传统数据平台不同的是，网络大数据海量的规模、多样的类型、快速的流动和动态的体系以及巨大的价值是大数据平台构建需要重点考虑的几个因素. 除此之外，数据的分类存储、数据平台的开放性、数据的智能处理以及数据平台与用户的交互都为网络大数据平台的建设带来前所未有的挑战. 网络大数据平台处理的数据类型是多种多样的. 根据数据类型的不同，网络大数据平台可以分为不同的类型，比如本体数据平台、企业日常事务数据平台、流数据平台、电子商务数据平台等等. 目前这些平台的搭建已经具有了一些有代表性的工作. 如 Google 公司的 Freebase①、微软公司的 Probase②[68]、国内著名的中文信息结构库——知网（Hownet）③等. 在商用数据平台方面，IBM 公司的 Infosphere 大数据分析平台④、天睿公司的 Teradata 统一数据环境⑤以及由国内天猫、阿里云、万网联合推出的国内首个电商云工作平台聚石塔⑥是 3 个典型的数据平台.

6.2　网络大数据下的高端数据分析

一个优秀的综合大数据处理平台不但可以为企业的决策和个人的生活提供服务，甚至还可以为国家政策的制定提供支持. 首先，依托大数据平台，国家可以分析各实体和产业之间的关联关系，从而了解行业发展的趋势，找到影响产业发展的关键性因素，统筹规划资金、人才、技术的良性流动与优化配置. 其次，大数据平台可以为企业提供巨大的商业价值. 企业分析人员可以分析多种多样的内容. 譬如，分析顾客偏好及顾客群体，对群体进行细分并量体

① http://www.freebase.com/
② http://research.microsoft.com/en-us/projects/probase/
③ http://www.keenage.com/
④ http://www.ibm.com/software/data/infosphere
⑤ http://www.teradata.com.cn/
⑥ http://cloud.tmall.com/index.htm

裁衣般地采取独特的行动;分析具有代表性的客户群体,采取有针对性的营销策略,进行病毒式营销和模式推广;运用大数据模拟实境,发掘新的需求和提高投入的回报率,进行商业模式、产品和服务的创新等.再次,大数据平台还可以为个人的日常生活带来诸多便利.建立在大数据平台下的互联网产业,将深加工的信息和数据主动推送给目标用户,便于用户结合自身喜好选择感兴趣的模式、产品和搭配方式.除此之外,用户还可以从大数据平台中获取更有价值的知识.通过本体知识平台,用户可以分析知识的来源、演化过程、分析知识间的因果关系、知识本身的歧义性和模糊性,更好地理解和关联知识.

6.3 网络大数据的应用

网络大数据平台在舆情监控、模式和关键字搜索、数据工程、情报分析、市场营销、医药卫生等领域具有重要的应用.举例来说,大数据平台的出现在搜索引擎中的应用是使得搜索引擎对数据的深入加工和处理变成现实,能够更好地理解用户的搜索意图.用户可以不用自己去筛选信息,而是由搜索引擎根据其搜索历史及个人偏好将有价值的信息呈现给用户.又如,网络大数据平台催生了很多面向程序员与数据科学家的工具(如 Karmasphere 和 Datameer),使得程序员将数据而非业务逻辑作为程序的主要实体,编写出更简短的程序,更清晰地表达对数据所做的处理.可以预见,大数据平台正在以一种前所未有的方式改变着各行各业,对大数据平台的应用能够更好地帮助人们获取信息并对信息进行更高效地处理和应用.

7 研究展望

当前在上述几个方向的研究工作都面临着网络大数据带来的新问题,也意味着每个方向都有不少的挑战.展望未来,面对网络大数据,以下几个方面的研究将是问题的核心.

网络大数据的复杂性度量.网络大数据使人们处理计算问题时获得了前所未有的大规模样本,但同时网络大数据也呈现出前所未有的复杂特征,不得不面对更加复杂的数据对象,其典型的特性是类型和模式多样、关联关系繁杂、质量良莠不齐.网络大数据内在的复杂性使得数据的感知、表达、理解和计算等多个环节面临着巨大的挑战,导致了传统全量数据计算模式下时空维度上计算复杂度的激增,很多传统的数据分析与挖掘任务如检索、主题发现、语义和情感分析等变得异常困难.然而目前,人们对网络大数据复杂性及其背后的物理意义缺乏理解,对网络大数据的分布与协作关联等规律认识不足,对大数据的复杂性和计算复杂性的内在联系缺乏深刻理解,加上缺少面向领域的大数据处理知识,极大地制约了人们对大数据高效计算模型和方法的设计能力.有鉴于此,如何量化定义大数据复杂性的本质特征及其外在度量指标,进而研究网络数据复杂性的内在机理是个重要的研究问题.

数据计算需要新模式与新范式.网络大数据的诸多突出特性使得传统的数据分析、数据挖掘、数据处理的方式方法都不再适用.因此,面对网络大数据,我们需要有数据密集型计算的基本模式和新型的计算范式,需要提出数据计算的效率评估方法等基本理论.由于数据体量太大,甚至有的数据本身就以分布式的形式存在,难以集中起来处理,因此对于网络大数据的计算需要从中心化的、自顶向下的模式转为去中心化的、自底向上、自组织的计算模式.而且,网络大数据来自于数量众多的网络用户.由于人为因素的随机性,网络大数据常常具有很高的噪声,同时也富含着冗余数据、甚至是垃圾数据.因此,面对网络大数据,去芜存精、化繁为简可能是必要的处理范式之一.另外,面对网络大数据将形成基于数据的智能,我们可能需要寻找类似"数据的体量+简单的逻辑"的方法去解决复杂问题.

新型的 IT 基础架构.网络大数据对于系统,不管是存储系统、传输系统还是计算系统都提出了很多苛刻的要求,现有的数据中心技术很难满足网络大数据的需求.因此,需要考虑对整个 IT 架构进行革命性的重构.而存储能力的增长远远赶不上数据的增长,因此设计最合理的分层存储架构,不仅满足 scale-up 式的可扩展性,而且还能满足 scale-out 式的可扩展性,已成为 IT 系统的关键.在大数据时代,IT 系统需要从数据围着处理器转改变为处理能力围着数据转,将计算推送给数据,而不是将数据推送给计算.此外,网络大数据平台(包括计算平台、传输平台、存储平台等)是网络大数据技术链条中的瓶颈,特别是网络大数据的高速传输,需要革命性的新技术.

数据的安全和隐私问题.数据有价值,有价值就可能产生争夺和侵害.只要有数据,就必然存在安全与隐私的问题.随着数据的增多,网络大数据面临着重大的风险和威胁,需要遵守更多更合理的规定,

而传统的数据保护方法无法满足这一要求.因此，面对网络大数据的安全与隐私保护，有大量的问题急需得到解决，具体包括：数据计算伦理学、数据密码学、分布式编程框架中的安全计算、远程数据计算的可信任度、数据存储和日志管理的安全性、基于隐私和商业利益保护的数据挖掘与分析、强制的访问控制和安全通信、多粒度访问控制以及数据来源和数据通道的可信等.

8 总　结

“人、机、物”三元世界融合的网络空间(Cyberspace)中的网络大数据存在数据规模巨大、数据关联复杂、数据状态演变等显著特征.其规模和复杂度的增长远远超出了符合摩尔定律增长的机器处理和计算能力.网络大数据带来了宝贵机遇，同时也存在着巨大挑战.本文从网络大数据的复杂性、不确定性和涌现性3个方面展开讨论，详细分析了这些特性给网络大数据的深度分析和价值利用带来的影响.本文梳理了网络大数据研究体系，从网络空间感知与数据表示、网络大数据存储与管理体系、网络数据挖掘和社会计算以及网络大数据平台系统与应用4个方面回顾了相关领域的新近发展，探讨了网络大数据研究方向和所面临的挑战，并展望了未来的主要研究方向.总之，与传统研究工作相比，网络大数据在各个层面的差异都非常显著.尽管目前已经有一些探索性的研究工作，但是总体上来说，网络大数据的研究还很年轻，尚有诸多问题亟待解决.

致　谢　本文的部分观点来自于香山科学会议第424次学术讨论会以及中国计算机学会大数据专家委员会针对大数据与网络大数据的深入讨论，本文的撰写还得到了孙晓明、郭嘉丰、沈华伟、兰艳艳等中国科学院计算技术研究所同事的大力支持，作者对相关专家与学者一并表示衷心的感谢！

参　考　文　献

[1] Li Guo-Jie, Cheng Xue-Qi. Research status and scientific thinking of big data. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647-657(in Chinese)
(李国杰，程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考. 中国科学院院刊，2012，27(6)：647-657)

[2] Big data. Nature, 2008, 455(7209): 1-136

[3] Dealing with data. Science, 2011, 331(6018): 639-806

[4] Holland J. Emergence: From Chaos to Order. Redwood City, California: Addison-Wesley, 1997

[5] Anthony J G Hey. The Fourth Paradigm: Data-intensive Scientific Discovery. Microsoft Research, 2009

[6] Phan X H, Nguyen L M, Horiguchi S. Learning to classify short and sparse text & Web with hidden topics from large-scale data collections//Proceedings of the 17th International Conference on World Wide Web. Beijing, China, 2008: 91-100

[7] Sahami M, Heilman T D. A web-based kernel function for measuring the similarity of short text snippets//Proceedings of the 15th International Conference on World Wide Web. Edinburgh, Scotland, 2006: 377-386

[8] Efron M, Organisciak P, Fenlon K. Improving retrieval of short texts through document expansion//Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. Portland, OR, USA, 2012: 911-920

[9] Hong L, Ahmed A, Gurumurthy S, Smola A J, Tsioutsiouliklis K. Discovering geographical topics in the twitter stream//Proceedings of the 21st International Conference on World Wide Web (WWW 2012). Lyon, France, 2012: 769-778

[10] Pozdnoukhov A, Kaiser C. Space-time dynamics of topics in streaming text//Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks. Chicago, IL, USA, 2011: 1-8

[11] Sun Yizhou, Norick Brandon, Han Jiawei, Yan Xifeng, Yu Philip S, Yu Xiao. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 1348-1356

[12] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York: Springer, 2009

[13] Meinshausen N, Yu B. Lasso-type recovery of sparse representations for high-dimensional data. The Annals of Statistics, 2009, 37(1): 246-270

[14] Zhou Ao-Ying, Jin Che-Qing, Wang Guo-Ren, Li Jian-Zhong. A survey on the management of uncertain data. Chinese Journal of Computers, 2009, 32(1): 1-16(in Chinese)
(周傲英，金澈清，王国仁，李建中. 不确定性数据管理技术研究综述. 计算机学报，2009，32(1)：1-16)

[15] Abiteboul S, Kanellakis P C, Grahne G. On the representation and querying of sets of possible worlds. Theoretical Computer Science, 1991, 78(1): 158-187

[16] Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques — Adaptive Computation and Machine Learning. Cambridge, MA: The MIT Press, 2009

[17] Aggarwal C C. Managing and Mining Uncertain Data. Berlin: Springer Publishing Company, Incorporated, 2009

[18] Wang Quan, Xu Jun, Li Hang, Craswell Nick. Regularized latent semantic indexing//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'11). Beijing, China, 2011: 685-694

[19] Mackey L, Talwalkar A, Jordan M I. Divide-and-conquer matrix factorization//Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS). Granada, Spain, 2011: 1134-1142

[20] Gershman S, Blei D. A tutorial on Bayesian nonparametric models. Journal of Mathematical Psychology, 2012, 56(1): 1-12

[21] Kulis B, Jordan M I. Revisiting k-means: New algorithms via Bayesian nonparametrics//Proceedings of the 29th International Conference on Machine Learning (ICML). Edinburgh, UK, 2012

[22] Yaneer Bar-Yam. A mathematical theory of strong emergence using multiscale variety. Complexity, 2004, 9(6): 15-24

[23] Bedau Mark A. Weak emergence. Noûs, 1997, 31(s11): 375-399

[24] Chalmers David J. Strong and Weak Emergence. Oxford: Oxford University Press, 2002

[25] Henrya Adam Douglas, Prałat Paweł, Zhangvol Cun-Quan. Emergence of segregation in evolving social networks. Proceedings of the National Academy of Sciences, 2011, 108(21): 8605-8610

[26] Bergman M K. White paper: The Deep Web: Surfacing hidden value. Journal of Electronic Publishing, 2001, 7(1). DOI: http://dx.doi.org/10.3998/3336451.0007.104

[27] Florescu D, Levy A, Mendelzon A. Database techniques for the World-Wide-Web: A survey. SIGMOD Record, 1998, 27(3): 59-74

[28] Fan Wenfei. Data quality: Theory and practice//Proceedings of the 2012 International Conference on Web-Age Information Management (WAIM'12). Harbin, China, 2012: 1-16

[29] Fan Wenfei, Geerts Floris. Foundations of data quality management. Synthesis Lectures on Data Management, 2012, 4(5): 1-217

[30] Fan Wenfei. Dependencies revisited for improving data quality//Proceedings of the 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS'08). Vancuver, Canada, 2008: 159-170

[31] Dasgupta A, Das G, Mannila H. A random walk approach to sampling hidden databases//Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. Beijing, China, 2007: 629-640

[32] Liu Wei, Meng Xiao-Feng, Ling Yan-Yan. A graph-based approach for Web database sampling. Journal of Software, 2008, 19(2): 179-193(in Chinese)
(刘伟, 孟小峰, 凌妍妍. 一种基于图模型的 Web 数据库采样方法. 软件学报, 2008, 19(2): 179-193)

[33] Wang R Y, Ben H B, Madnick S E. Data quality requirements analysis and modeling//Proceedings of the 9th International Conference on Data Engineering. Vienna, Austria, 1993: 670-677

[34] Galhardas H, Florescu D, Shasha D, Simon E. AJAX: An extensible data cleaning tool. ACM SIGMOD Record, 2000, 29(2): 590

[35] Guo Zhi-Mao, Zhou Ao-Ying. Research on data quality and data cleaning: A survey. Journal of Software, 2002, 13(11): 2076-2082(in Chinese)
(郭志懋, 周傲英. 数据质量和数据清洗研究综述. 软件学报, 2002, 13(11): 2076-2082)

[36] Hernandez M A, Stolfo S J. Real-world data is dirty: Data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery, 1998, 2(1): 9-37

[37] Abiteboul S, Cluet S, Milo T, Mogilevsky P, Simeon J, Zohar S. Tools for data translation and integration. IEEE Data Engineering Bulletin, 1999, 22(1): 3-8

[38] Milo T, Zohar S. Using schema matching to simplify heterogeneous data translation//Proceedings of the 24th International Conference on Very Large Data Bases. New York, NY, USA, 1998: 122-133

[39] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990, 41(6): 391-407

[40] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3(1): 993-1022

[41] Guo Jiafeng, Xu Fu, Cheng Xueqi, et al. Named entity recognition in query//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09). Boston, MA, USA, 2009: 267-274

[42] Chang F, Dean J, Ghemawat S, Hsieh W C, et al. A distributed storage system for structures data//Proceedings of the 7th Symposium on Operating Systems Design and Implementation. Seattle, WA, USA, 2006: 205-218

[43] Abadi D, Madden S, Hachem N. Column-Stores vs. Row-Stores: How different are they really ?//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. Vancouver, BC, Canada, 2008: 967-980

[44] He Yongqiang, Lee Rubao, Huai Yin, Shao Zheng, Jain N, Zhang Xiaodong, Xu Zhiwei. RCFile: A fast and space-efficient data placement structure in MapReduce-based warehouse systems//Proceedings of the 2011 IEEE 27th International Conference on Data Engineering (ICDE). Hannover, Germany, 2011: 1199-1208

[45] Zou Y Q, Liu J, Wang S C, Zha L, Xu Z W. CCIndex: A complemental clustering index on distributed ordered tables for multi-dimensional range queries//Proceedings of the Network and Parallel Computing. Zhengzhou, China, 2010: 247-261

[46] Gao Ming, Jin Che-Qing, Wang Xiao-Ling, Tian Xiu-Xia, Zhou Ao-Ying. A survey on management of data provenance. Chinese Journal of Computers, 2010, 33(3): 373-389 (in Chinese)
(高明，金澈清，王晓玲，田秀霞，周傲英. 数据世系管理技术研究综述. 计算机学报，2010，33(3)：373-389)

[47] Buneman P, Khanna S, Tan Wang-Chiew. Data provenance: Some basic issues//Proceedings of the Software Technology and Theoretical Science. New Delhi, India, 2000: 87-93

[48] Tan W C. Provenance in databases: Past, current, and future. IEEE Data Engineering Bulletin, 2007, 30(4): 3-12

[49] Gong Xue-Qing, Jin Che-Qing, Wang Xiao-Ling, Zhang Rong, Zhou Ao-Ying. Data-intensive science and engineering: Requirements and challenges. Chinese Journal of Computers, 2012, 35(8): 1563-1578(in Chinese)
(宫学庆，金澈清，王晓玲，张蓉，周傲英. 数据密集型科学与工程：需求和挑战. 计算机学报，2012，35(8)：1563-1578)

[50] Li P, Burges C, Wu Q. McRank: Learning to rank using multiple classification and gradient boosting//Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS'07). Vancouver, BC, Canada, 2007, 19: 845-852

[51] Freund Y, Iyer R, Schapire R E, et al. An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research, 2003, 4: 933-969

[52] Burges C, Shaked T, Renshaw E, et al. Learning to rank using gradient descent//Proceedings of the 22nd International Conference on Machine Learning (ICML'05). Bonn, Germany, 2005: 89-96

[53] Cao Zhe, Qin Tao, Liu Tie-Yan, et al. Learning to rank: From pairwise approach to listwise approach//Proceedings of the International Conference on Machine Learning (ICML'07). Corvallis, OR, USA, 2007: 129-136

[54] Xu Jun, Li Hang. AdaRank: A boosting algorithm for information retrieval//Proceedings of the 31st International ACM SIGIR Conference (SIGIR'07). Amsterdam, 2007: 391-398

[55] Yue Y, Finley T, Radlinski F. A support vector method for optimizing average precision//Proceedings of the 31st International ACM SIGIR Conference (SIGIR'07). Amsterdam, 2007: 271-278

[56] Cheng Xue-Qi, Guo Jia-Feng, Jin Xiao-Long. A retrospective of Web information retrieval and mining. Journal of Chinese Information Processing, 2011, 25(6): 111-117(in Chinese)
(程学旗，郭嘉丰，靳小龙. 网络信息的检索与挖掘回顾. 中文信息学报，2011，25(6)：111-117)

[57] Yangarber R, Grishman R. NYU: Description of the Proteus/PET system as used for MUC-7//Proceedings of the 7th Message Understanding Conference (MUC'98). Fairfax, VA, 1998

[58] Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction. Journal of Machine Learning Research, 2003, 3(1): 1083-1106

[59] Getoor L, Taskar B. Introduction to Statistical Relational Learning. Cambridge, MA: The MIT Press, 2007

[60] Shen H W, Cheng X Q, Guo J F. Quantifying and identifying the overlapping community structure in networks. Journal of Statistical Mechanics: Theory and Experiment, 2009, (7): P07042

[61] Shen Hua-Wei, Jin Xiao-Long, Ren Fu-Xin, Cheng Xue-Qi. Analysis on social media. Communication of the CCF, 2012, 8(4): 32-36(in Chinese)
(沈华伟，靳小龙，任福新，程学旗. 面向社会媒体的舆情分析. 中国计算机学会通讯，2012，8(4)：32-36)

[62] Girvan M, Newman M E J. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826

[63] Newman M E J, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2): 026113

[64] Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005, 435(7043): 814-818

[65] Shen H W, Cheng X Q, Cai K, Hu M B. Detect overlapping and hierarchical community structure in networks. Physica A, 2009, 388(8): 1706-1712

[66] Barabási A L, Albert R. Emergence of scaling in random networks. Science, 1999, 286(5439): 509-512

[67] Palla G, Barabási A L, Vicsek T. Quantifying the social group evolution. Nature, 2007, 446(7136): 664-667

[68] Wu Wentao, Li Hongsong, Wang Haixun, Zhu Kenny. Probase: A probabilistic taxonomy for text understanding//Proceedings of the 2012 International Conference on Management of Data (SIGMOD). Scottsdale, AZ, USA, 2012: 481-492

WANG Yuan-Zhuo, born in 1978, Ph. D., associate professor. His current research interests include social computing, network security analysis, stochastic game model, etc.

JIN Xiao-Long, born in 1976, Ph. D., associate professor, Ph. D. Supervisor. His research interests include social computing, network performance modelling and evaluation.

CHENG Xue-Qi, born in 1971, Ph. D., professor, Ph. D. supervisor. His research interests include network science, network security analysis, Web search & data mining.

Background

Traditionally, massive data are mostly produced in scientific fields such as astronomy, meteorology, genomics physics, biology, and environmental research. Due to the rapid development of IT technology and the consequent decrease of the cost on collecting and storing data, massive data have been being generated from almost every industry and sector as well as governmental department, including retail, finance, banking, security, audit, electric power, healthcare, to name a few. Network big data are the massive data generated by interaction and fusion of the ternary human-machine-thing universe in the Cyberspace, which includes tons of user generated contents, log files, deep web data, etc. Network big data have attracted extensive interests from both academia and industry due to the potential big social, commercial, and scientific value.

Evidently, network big data have been not only changing the way in which people live and work, but also reforming the mode that enterprises run. The famous McKinsey Global Institute regards big data as 'the next frontier for innovation, competition, and productivity'. Nature and Science have published special issues in 2008 and 2011, respectively, to discuss the unprecedented opportunities that big data bring to us. Moreover, the US government announced in 2012 a "Big Data Research and Development Initiative" aiming to greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of big data. Because of this background, an academic symposium of the Xiangshan Science Conferences was organized last year in Beijing to discuss the issues of the challenges, theoretical foundation, and ecosystem of network big data. Later on, the China Computer Federation (CCF) founded the CCF Task Force on Big Data (CCF TFBD) to investigate and study the core scientific and technological issues of network big data.

Although network big data bring us unparalleled opportunities, they, however, also pose many grand challenges to us. Particularly, due to the features of big data such as sea-scale volume and heterogeneous formats, the existing theory and technology of data processing cannot efficiently and effectively cope with network big data. As a consequence, it calls for a specialized discipline to explore the common laws of network big data, study fundamental theory and essential approaches to qualitatively or quantitatively handling network big data, and eventually build solid foundation for developing new theory, techniques, and methods for big data processing. This paper summarizes the research issues and present status of network big data and looks ahead to the development trends.